DOCUMENT RESUME

ED 182 306 SP 015 616

AUTHOR Capie, William: And Others
TITLE The Objectivity of Classroom Observation Scales.
SPONS AGENCY Georgia Univ., Athens.
PUB DATE Nov 78
NOTE 9p.; Paper presented at the Annual Meeting of the Southeastern Association for the Education of Teachers of Science (New Orleans, Louisiana, November 29, 1978)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Behavior Rating Scales; *Classroom Observation Techniques; *Interaction Process Analysis; *Measurement Instruments; Measurement Techniques; Observation; Teacher Behavior; *Teacher Evaluation

ABSTRACT

The purpose of this study was to determine the degree to which inter-rater agreement in classroom observation of teachers is affected by using a system where each scale point is defined by a descriptor particularly created for the quality being rated. Thirty-three subjects enrolled in two Teacher Performance Assessment Instruments training sessions were involved in the study. About two-thirds of the trained observers were able to choose the same scale point when descriptors were available. Rates of agreement were increased from about 45 percent to about 70 percent when descriptors were available and observers trained in their use. Differences in mean ratings suggested that different qualities were assessed when no descriptors were used. (JD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED182306

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

THE OBJECTIVITY OF CLASSROOM OBSERVATION SCALES*

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Kathy Miller

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

William Capie
The University of Georgia

Joan Kolok
The University of the South

L. Antonette Hale
Florida State University

Paper presented at the
Annual Meeting of the
Southeastern Association for the Education
of Teachers of Science
New Orleans November 29, 1978

*The research reported in this paper was partially supported by the Teacher Assessment Project, University of Georgia, Athens, GA

SP015616

THE OBJECTIVITY OF CLASSROOM OBSERVATION SCALES

Background

Classroom observation scales are widely used to evaluate teachers and to describe teaching performances. Two item formats are common. In the simplest case a set of standard response options is used for a variety of item stems. For example, a list of twenty traits is provided for observers to rate each one poor, fair, adequate, good, outstanding; 1-5; always, almost always, etc.. Occasionally only the endpoints of the scale are defined with the middle points undefined and marked according to the rater's judgement.

The Georgia Teacher Performance Assessment Instruments (TPAI) (Okey, et al, 1978) have been developed using another approach. There is a set of approximately 50 behavioral indicator statements each of which is rated on a 1-5 scale. However, in an attempt to increase reliability, each of the 5 scale points is defined by a one or two sentence descriptor so that its meaning is more explicit. The observer must pick the descriptor statement which is most characteristic of the teaching performance. A sample indicator is shown in Figure 1.

Attends to routine tasks

1. The teacher does not attend to routine tasks.
2. The teacher attends to routine tasks in a disruptive or inefficient manner (e.g., learners need special permission for many routine tasks).
3. The teacher anticipates routine tasks and attends to them efficiently. Not all work stops for everyone while some are being accommodated.
4. Routine tasks are handled smoothly--the teacher and learners work together efficiently.

5. In addition to #4, learners are responsible for various dimensions of the tasks (e.g., distributing materials, picking up work area, etc.).

Figure 1
A Sample Indicator from the TPAI

Problem Statement

The purpose of this study was to determine the degree to which inter-rater agreement is affected by using a system where each scale point is defined by a descriptor particularly created for the quality being rated as in the TPAI.

Procedures

Thirty-three subjects enrolled in two TPAI training sessions were involved in the study. Data were collected on two occasions. As part of the training orientation, subjects watched a videotape of a 30 minute classroom lesson and rated the teaching performance. Specially modified versions of two TPAI's were used for this purpose. Twenty-two behavioral indicator statements dealing with Classroom Procedures (CP) and 9 statements reflecting Interpersonal Skills (IS) were presented. Unaware of the descriptors, trainees were told to rate each performance on a 1-5 scale with 5 being best and 1 the weakest demonstration of a performance. One group rated a videotape called "An Eighth Grade Language Arts Teacher" and the other rated "A Third Grade Science Teacher". (Capie, et al, 1978). These videotapes were developed for training raters to use the TPAI.

After approximately 8 hours of training in the use of the instruments, the second tape was rated using the TPAI with the descriptors. Training included presenting descriptors, defining terms, citing examples and practicing rating additional tapes with feedback. This procedure is described in the TPAI training program (Capie, et al, 1978). The data generated

on these two rating occasions were then analyzed for interrater agreement.

Group 1 n=12	Rate 8th Grade Language Arts without descriptors	Undergo Training	Rate 3rd Grade Science with descriptors
Group 2 n=21	Rate 3rd Grade Science without descriptors	Undergo Training	Rate 8th Grade Language Arts with descriptors

Figure 2.
Treatment and Data Collecting Design

Data

For each indicator, 22 on CP and 9 on IS, the number of individuals who selected the mode was determined. This number was compared to the group size to calculate the percentage of agreement. For example, if half the participants chose the modal response for an indicator, its percentage of agreement was 50. The mean percentage of agreement for all indicators on each instrument was computed for each rating occasion. Since each training group viewed a different lesson with and without descriptors, data were collapsed over both pre- and over both post-tapes. Using both tapes as part of the "with" and "without descriptor" ratings was an attempt to control the effect of differences in the ambiguity of the tapes.

Analysis

A X^2 test for the difference between the proportions was calculated comparing the mean rates of agreement with and without descriptors for both instruments.

Results

For the Classroom Procedures the mean rates of agreemnt were .46

without descriptors and .65 with the descriptors. On the Interpersonal Skills the mean agreement values were .44 and .70. Complete data are shown in Table 1. Chi square testing indicated both sets of proportions were different beyond the .01 level (with 1 degree of freedom Interpersonal Skills X^2 = 40.7; Classroom Procedures X^2 = 53.1).

	Classroom* Procedures	Interpersonal** Skills
Mean Rate of Agreement With No Descriptors	.46	.44
Mean Rate of Agreement With Descriptors	.65	.70

*$X^2_{2\,1df}$ = 53.1, differences significant beyond .01 level.

**$X^2_{1\,df}$ = 40.7, differences significant beyond .01 level.

***effective N was 594 for IS and 1452 for CP since 33 subjects rated 2 lessons and 9 and 22 indicators respectively.

TABLE 1
Comparing Mean Level of Agreement With and Without Descriptors

Discussion

The results of this study are encouraging to the instrument developers for two reasons. First, the levels of agreement after training in use of the descriptors are high for these moderate to high inference level ratings, and individuals who didn't choose the modal response tended to converge on the same second choice. A second cause for satisfaction is the evidence that the descriptors do indeed reduce amibuity in the manner that had been expected. Agreement was increased 41% on the CP and 59% on the IS.

Levels of agreement with and without descriptors should be of interest to individuals contemplating the use of similar scales. The sixty-five to seventy percent agreement may represent the maximum possible agreement on items of this type since a great deal of development effort has gone into

the instruments and the training associated with them. Using the training as part of the "with" descriptor treatment was done to maximize effects and because it is part of the system used in Georgia. The 45 percent agreement in the pre-ratings may not represent the floor, however, since considerable care has gone into the development of the behavioral indicator statements, as well.

Superficial inspection of the data may suggest that the original ratings were "just a little off". In fact, this was not so. When near-agreement (the mode plus one adjacent scale point) was considered, agreement was approximately 0.70 without descriptors, but in excess of 0.90 when descriptors were used. These data are summarized in Table 2.

	Classroom* Procedures	Interpersonal** Skills
Rate of Near-Agreement With No Descriptors	.72	.72
Rate of Near-Agreement With Descriptors	.92	.90

*X^2_{1df} = 98.1, differences significant beyond .01 level.

**X^2_{1df} = 30.7, differences significant beyond .01 level.

TABLE 2
Comparing Mean Levels of Near Agreement
With and Without Descriptors

One other interesting observation is that the mean ratings for the same lessons were substantially lower, .6 to .8 scale points, when comparing the initial ratings with ratings made with descriptors. These data suggest that different dimensions may be tapped when no descriptors are used. The ratings with descriptors are in line with standard profiles prepared by a panel of "experts" who rated the tapes. These data are shown in Table 3.

	Classroom Procedures		Interpersonal Skills	
	Science	Language Arts	Science	Language Arts
Mean With No Descriptors	3.34	3.98	3.94	3.59
Mean With Descriptors	3.98	4.04	4.71	4.36
Mean of Standard Ragings	4.09	4.02	4.47	4.16

TABLE 3
Mean Ratings on All Indicators of Two Lessons
With and Without Descriptors

In summary, on the average, about two-thirds of the trained observers were able to choose the same scale point when descriptors were available. Rates of agreement were increased from about 45 percent to about 70 percent when descriptors were available and observers trained in their use. Furthermore, differences in mean ratings suggested that different qualities were assessed when no descriptors were used.

Okey, J.; Capie, W.; and Johnson, C.E., Teacher Performance Assessment Instruments: Manual for Data Collectors, University of Georgia, Athens, 1978.

Capie, W., Anderson, S.J., Okey, J., and Johnson, C.E., "Classroom Teacher Performance: A Videotape Series", University of Georgia Athens, 1978.

Capie, W., Anderson, S.J., Okey, J., and Johnson, C.E., Teacher Performance Assessment Instruments: Conducting a Training Session, University of Georgia, 1978.